DOCUMENT RESUME

ED 219 396 TM 820 399

AUTHOR Klibanoff, Leonard S.
TITLE Implementation Factors and Bias in National Impact Data on Title I.
PUB DATE Mar 82
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Gains; Bias; Elementary Secondary Education; Evaluation Methods; *Federal Programs; *Program Evaluation; *Research Problems; State Departments of Education; *Statistical Bias
IDENTIFIERS Evaluation Problems; Normal Curve Equivalent Scores; *Title I Evaluation and Reporting System

ABSTRACT

An attempt was made to estimate possible bias in terms of Normal Curve Equivalent (NCE) gains of Elementary Secondary Education Act (ESEA) Title I projects. Twenty-two state Title I reports for 1978 were examined. The analyses presented were directed at two major sources of bias: bias due to errors in local implementation of models, and bias due to incomplete responses. Two subsets of five states each were identified where the State Education Agency (SEA) attempted to ameliorate implementation problems and where no such action was taken. Gains for the two subsets were compared. The numerical gains were large across the group of 22 states. The numerical gains were larger than the verbal gains, and lower grade gains were consistently greater than those in upper grades. Bias from errors in implementation was estimated at from four to ten NCEs. States without implementation controls showed consistently higher gains. Three additional sources of bias were non-response by SEAs, correctable errors, and further regression to the mean in properly implemented Model A evaluations. These sources of bias were identified, but not analyzed. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219396

IMPLEMENTATION FACTORS AND BIAS IN NATIONAL IMPACT DATA ON TITLE I

Leonard S. Klibanoff

RMC Research Corporation
Mountain View, CA 94040

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K Klibanoff, L.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 399

Paper presented at
The Annual Meeting of the
American Educational Research Association
March 19-23, 1982
New York, NY

IMPLEMENTATION FACTORS AND BIAS IN NATIONAL IMPACT DATA ON TITLE I

The 1977-78 school year (FY 78) was the first year of widespread use of the new USOE Title I evaluation and reporting system. Use of the system was voluntary at the State level and in many States at the district level. Some 23 States reported achievement gains in terms of Normal Curve Equivalents (NCEs) within their jurisdictions. Achievement data from 23 other State educational agencies (SEAs) were reported, but were not in NCEs or the reports were received after the deadline. Not all of the results in the 23 States are based entirely on one of the USOE models. Some reports contain a mixture of achievement data obtained through use of one of the models and data obtained through other evaluation designs and reported in metrics other than NCEs (e.g., percentiles or grade equivalents). Because of the inchoate nature of the implementation of the evaluation models, more than the usual amount of concern exists for the validity of the achievement gains reported by the States to USOE. The purpose of this paper is to provide some estimates of the size of the bias in the NCE gains.

An examination of 22 of the State reports available to RMC was undertaken for the purpose of defining or estimating some confidence interval around the reported gain. Because of the pilot nature of the FY 78 implementation of the models, an unknown, but potentially substantial, degree of bias is expected due to the voluntary use of the models within States. Only a handful of the States had a substantial majority of their LEAs implement one of the models. Hence, the biases attendant on differential responses (i.e., selection bias) are sure to be present in the data of nearly all States. Although we have made some estimates of the response bias, our major concern is not, however, directed at the selection or response effects, but rather at possible bias associated with improper or incomplete implementation of the models.

Method

We first examined the reports to determine what information they contained on the actual implementation of the evaluation models. In a number of cases (thirteen), little or no information is available on those implementation aspects that could affect the gain estimates. Few States report any particular SEA actions taken to "clean" the data to increase their accuracy. Although some of the reports contain no information on possible difficulties, others indicate some problems at the local level, still others indicate problems Statewide, a few indicate some useful steps taken at the State level to isolate clear cases of model violations prior to aggregating the results to the State level.

The data discussed in this paper are based primarily on the results of evaluations using Model A. Most States that identify which models were used indicate that it was the norm-referenced model. Where States did not indicate which models were used, we assumed that Model A generated the data. In one State we were able to identify the results by model and we included only the Model A results. Our focus on Model A stems from its being far and away the most frequently used model. In particular, we attempted to identify those cases where selection on the pretest score was evident. Unlike most of the other problems that can arise in implementing Model A, selection on the pretest for participation in Title I has a positive bias associated with it. Other implementation problems, for example, encountering test floors or ceilings or using tests with interpolated or extrapolated norms, can have compensating effects (i.e., both over and underestimates result) during the course of aggregation.

Using State reports alone does not allow a very detailed explication of the implementation problem that could affect the gain estimates. The reports differ substantially among themselves with respect to the amount and quality of information they contain on either SEA or LEA activities. A more precise and detailed study of implementation problems (or more generally model violations) would require an examination of the LEA reports and, perhaps, discussions with LEA personnel. In our review of the

reports we were less concerned with isolating the existence and frequency of possible violations of the implementation rules, than with identifying a subset of States that appear to have provided relatively "cleaner" data than most other States. Finding such a subset of States and comparing their results with those of other States provides at least some basis for assessing the degree of bias. Tabulating the frequency of occurrence of various errors in applying the models (even if it were possible to do so with only the State reports) during a pilot year would not provide such a basis. A small subset (five) of such States was identified. In these five States at least one of the following conditions prevailed: (1) cases selected on the pretest were eliminated from the aggregation; (2) high and low gain estimates were provided; or (3) where selection on the pretest was prevalent, it was not the sole selection factor. Similarly, we identified a subset of five States where no SEA action had been taken to ameliorate implementation problems. The reports for the States in this subset provided little or no information about quality control activities. Telephone conversations with the relevant Technical Assistance Center (TAC) for each State confirmed that there indeed had been implementation problems and/or no quality control at the SEA level.

Results

Estimates of Bias Due to Model Violation

Table 1 presents gains aggregated across States and grades 2-12 and separately for grade bands 2-4, 5-6, and 7-12. There are three noteworthy points in the data of Table 1. First, the size of the gains are large. For example, the numerical gain over grades 2-12 is slightly more than one-third of the population standard deviation in NCE units (21.06 NCEs). Such a large gain for twenty-two States seems unlikely to be unbiased. Gains of this level of magnitude would not only be at variance with most research on the effects of Title I it would also represent a substantial shift in the Title I achievement distribution. To obtain such a large shift in the location of the mean implies that, assuming a normal distribution of gains, half of the Title I students are achieving gains at and

above 7 NCEs, the level considered by some as indicating educational significance. Second, the numerical gains are uniformly larger than the verbal gains. Third, gains made in the lower grades are consistently greater than those made in the higher grades. Both the differences between numerical and verbal gains and the trend over grade levels may result in part from differential selection on the pretest and differential test reliabilities by grade and subject matter. However, both sets of differences (between subject areas and over grades) may also be due in part to real differences in the effectiveness of numerical versus verbal projects and of projects serving younger rather than older students.

Table 1

Weighted NCE Gains Aggregated to the National Level (Model A)

	Verbal			Numerical		
Grade Band	Gain	S.D.	N*	Gain	S.D.	N*
2-12	6.44	2.78	508,713	7.14	3.25	192,509
2-4	7.80	3.06	278,811	8.60	3.48	87,979
5-6	5.38	1.90	136,892	6.61	3.40	63,991
7-12	3.91	2.26	92,405	4.70	2.65	39,881

*The number of students in the total of grades 2-12 is greater than the sum of the grade based N's due to inclusion of students in ungraded classrooms in the total.

Similar results are shown for the two subsets of SEAs described earlier. Group A consists of five States which appeared to have taken some ameliorative action to mitigate the effects of local problems in correctly implementing the evaluation models. Group B consists of five States that appeared not to have taken such action or which provided no information. We should point out that the SEAs in Group A are not those reporting the smallest aggregate gains, nor are those in Group B the States showing the largest gains. On balance, the two groups are roughly comparable in terms of the aggregate size of their Title I populations, about 202,000 in Group A and about 280,000 in Group B and in the percentage of students for whom they have achievement gains in NCEs (response

rate), about 58% in Group A and 44% in Group B. (Please see the following section for a description of how population size and response rate are estimated.) Although comparisons between the two groups of SEAs do not form a completely sound basis for accurately defining the bias in the aggregate data, they do provide some estimate of the extent of the bias. Given the lack of information in the State reports on this issue, much beyond suggestive estimates of the bias cannot be provided.

The results of the comparisons are shown in Table 2. The Group A SEAs show average gains ranging from 1.73 to 3.66 NCEs in verbal projects and 2.55 to 4.52 NCEs in numerical projects. The Group B SEAs show consistently higher gains (6.04 to 11.43 NCEs and 7.75 to 14.22 NCEs in verbal and numerical projects respectively). For the full set of grades 2-12 the SEAs in Group B reported gains of 6.7 and 8.6 NCEs greater than the Group A SEAs in verbal and numerical projects respectively.

Table 2

Weighted NCE Gains Aggregated Across States in Group A and Group B

Grade Band	Group A*			Group B		
	Gain	S.D.	N**	Gain	S.D.	N
			Verbal			
2-12	3.12	1.77	117,435	9.82	1.64	124,351
2-4	3.66	1.87	54,594	11.43	1.13	79,759
5-6	3.55	.94	31,774	7.50	1.16	29,292
7-12	1.73	2.23	31,067	6.04	1.64	14,695
			Numerical			
2-12	3.64	1.24	62,411	12.24	1.91	21,105
2-4	4.52	.75	25,395	14.22	2.63	8,928
5-6	3.41	1.73	20,948	12.51	1.41	7,670
7-12	2.55	1.32	16,068	7.75	1.25	4,299

*See text for description of SEAs in Groups A and B.
**N refers to the number of students included in the aggregation.

The reported gains in both groups of States are higher in numerical achievement than in verbal and in the lower grades than in the higher

grades. The differences between the two groups are shown in Table 3. As indicated above, these differences, while suggestive of the bias associated with the quality of model implementation within States, cannot serve as point estimates of that bias. Because the differences between subject areas and among grade levels may be real, it is difficult to conclude firmly that the bias is larger in the lower grades and in numerical achievement than in the higher grades and in verbal achievement. It is likely that these results reflect both bias and real differences.

Table 3

Differences in NCE Gains Reported
by Two Groups of SEAs

	NCE Gain Differences	
Grade Band	Verbal	Numerical
2-12	6.70	8.60
2-4	7.77	9.70
5-6	3.95	9.10
7-12	4.31	5.20

Although the method involving subtracting the gains reported by one group of States from those reported by another group is an ad hoc procedure for estimating the bias in the national aggregate, it does indicate that a bias of from 4 to 10 NCEs may be reflected in the national results. Since the degree to which our classification of the States into the two groups captures differences in the quality of Model A implementation in those States is unknown, a more precise estimate beyond the range of 4 to 10 NCEs is not possible. We believe that the bulk of the bias results from selection on the pretest in Model A; however, it is likely that additional kinds of bias from other sources are also present. One such source may be the different response rates within the States in each group. Other sources of bias are discussed below.

Bias Associated with Response Rates

In order to estimate the bias due to non-response (self-selection on the part of districts in accordance with the voluntary nature of the FY 78

implementation as well as incomplete responses from participating districts) within the States, one needs to know at least the size of the Title I participating population, the response rates, and the correlation between response and NCE gains. Our means for approximating these data are discussed on the following paragraphs.

In estimating the response rate represented by the twenty-two available reports we were forced to use different measures of the participating population within various States. In some cases, the unduplicated total counts by grade level were used; in other cases the participant population figures were available by subject (service) area within grade levels; in still other cases only totals (not identified as duplicated or unduplicated) over all grades served were available. Because of the lack of consistent records, we attempted to arrive at a rough estimate of the Title I population in these twenty-two States. To do so we took the available counts for grades 2-12 (usually the unduplicated totals) and used that as the estimate of the population size. Because of the differing counts provided by the States and because our tabulations of those counts are probably underestimates, we have used the total number of students for whom NCE gains in reading were reported as the sample size for determining the estimate of the overall response rate.

Using the procedure described above, we estimate that the within-State response pooled across the twenty-two States is on the order of 42%. This probably represents a liberal estimate. By using tables developed by A. O. H. Roberts of the maximum bias due to non-response, assuming a normal distribution, we can estimate the size of the non-response bias as a function of the correlation between response and NCE gain within the States. Roberts' tables are based on the formula:

$$\text{Bias} = r\,\frac{Y}{P}$$

where

r = correlation between response and gain.

Y = height of the normal deviate at a cutoff corresponding to P.

P = rate of response.

Since that correlation is unknown, we provide estimates of the bias for various values of the correlation in Table 4. Also shown are the bias estimates for selected response rates ranging from 10 to 99%.

Table 4

Possible Bias in NCE Gains Due to Non-Response as a Function of Various Correlations between Response and Achievement Gains

Response Rate (%)	Correlation of Response and Gain									
	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
10	3.71	7.41	11.12	14.83	18.54	22.24	25.95	29.66	33.36	37.07
20	2.95	5.90	8.85	11.79	14.74	17.69	20.64	23.59	26.54	29.48
30	2.44	4.89	7.33	9.77	12.21	14.66	17.10	19.54	21.99	24.43
40	2.04	4.09	6.13	8.17	10.21	12.26	14.30	16.34	18.39	20.43
50	1.68	3.37	5.05	6.74	8.42	10.11	11.79	13.48	15.16	16.85
60	1.35	2.70	4.04	5.39	6.74	8.09	9.43	10.78	12.13	13.48
70	1.05	2.11	3.16	4.21	5.27	6.32	7.37	8.42	9.48	10.53
80	.74	1.47	2.21	2.95	3.69	4.42	5.16	5.90	6.63	7.37
90	.42	.84	1.26	1.68	2.11	2.53	2.95	3.37	3.79	4.21
99	.06	.13	.19	.25	.32	.38	.44	.51	.57	.63

Given our estimated response rate of 42% and assuming a correlation between response and gain of .10 to .40, the bias due to non-response within States falls between 1.96 and 7.83 NCEs. Since we have little *a priori* basis for choosing any given correlation (and since our estimate of the response rate is rather rough), a more precise estimate of the bias due to non-response is probably not attainable.

Additional Sources of Bias

There are three additional sources of bias that may be affecting the gain estimates. First, although we have attempted to estimate the range of bias that might be caused by non-response within States, the gains may also be reflecting some bias from non-response by SEAs as well. Since our analyses are based on a subset of SEAs, additional bias may exist.

The second additional source, not dealt with directly in our analyses, is the "correctable" error made at the local level. Correctable errors are defined as calculation or conversion errors (e.g., using the wrong norms table) that could be corrected at the raw data level. Crane and Maye (1980) showed a net positive bias on the order of one-quarter to three-quarters of an NCE in one State's data.

A third source of bias is one that appears to be associated even with correct implementations of Model A. Roberts (1980) reports that some additional regression to the mean occurs in Model A in the absence of selection on the pretest. The magnitude of this bias is about one-half to one NCE. Since the primary implementation source of bias is selection on the pretest, this further regression to the mean is probably largely reflected in the estimate of bias given earlier.

Summary and Conclusions

The analyses presented were directed at two major sources of bias. Namely, bias due to errors in the local implementation of the models and bias due to incomplete responses by LEAs within the 22 states. Three additional sources of bias were identified but not analyzed in this report. These are non-response by SEAs, "correctable" errors, and further regression to the mean in properly implemented Model A evaluations.

The extent of the bias caused by violations of the rules for implementing Model A was estimated to be in the range of 4 to 10 NCEs. This estimate was obtained by comparing the gains reported by two groups of States. Because the method used is not entirely satisfactory, the estimate is best viewed as suggestive of the true magnitude of the bias. It should serve, however, to facilitate the interpretation of these gains in relation to future years gains that will not be as subject to implementation errors. As users become more expert in implementing the system, the extent of bias caused by improper implementation should be dramatically reduced.

Our analysis of the bias caused by non-response within the 22 states indicates that the bias may be substantial. Assuming likely levels of the correlation betwen response and gain, we estimate a bias of about 2 to 8 NCEs due to the fact that not all LEAs in the States that were studied submitted NCE gains. If using the models and reporting their results are independent of achievement gains, then, of course, there would be no bias. However, it seems reasonable to assume at least a weak correlation and, hence, a bias at the lower end of the indicated range. As the models come into universal use, bias due to non-response should diminish rapidly.

It is important to note that the two major sources of bias (implementation errors and non-response) are not necessarily independent and additive. Merely adding the bias estimates together to get an estimate of the total bias is not appropriate. The tendency not to implement one of the models or not to report the results to the SEA may be correlated with the tendency to encounter implementation difficulties or to implement the models improperly. Treating the estimates of bias as uncorrelated and additive would be imprudent.

In sum, it would appear that the bias in the 1977-78 gain estimates is on the order of 4 to 10 NCEs. It is possible, however, that the bias could be somewhat larger, say 6 to 12 NCEs. Considering that many States were implementing the Title I evaluation and reporting system for the first time in 1977-78, bias of the indicated magnitude is neither surprising nor especially disconcerting. We should, however, be aware of the bias as we interpret gains from future years reports.

REFERENCES

Crane, L. R., & Maye, R. O. Effects of correctable errors on Title I NCE gain estimates and implications for statistical quality control. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April 1980.

Roberts, A. O. H. Regression to the mean and the regression effect bias. Mountain View, CA: RMC Research Corporation, 1980.